A Free Energy Principle for the Brain (Japanese Translation)

Abstract

ヘルムホルツの知覚に関する考え方を現代の理論で定式化すると、驚くほど広範な神経生物学的事実を説明できる知覚推論と学習のモデルが得られます。統計物理学の構成要素を用いることで、感覚入力の原因を推論する問題と、その生成の因果構造を学習する問題が、全く同じ原理で解決できます。さらに、推論と学習は生物学的に妥当な方法で進行できます。結果として得られるスキームは、経験ベイズ（Empirical Bayes）と、感覚入力がどのように引き起こされるかに関する階層モデルに基づいています。階層モデルの使用により、脳は動的かつ文脈に依存した方法で事前期待（prior expectations）を構築することができます。このスキームは、皮質の組織化と反応の多くの側面を理解するための原理的な方法を提供します。

本稿では、これらの知覚プロセスが、自由エネルギー原理に従うシステムの創発的行動の一側面であるにすぎないことを示します。ここで考慮される自由エネルギーは、システムに作用する環境量の確率分布と、その構成によって符号化された任意の分布との間の差を測定します。システムは、環境をサンプリングする方法に影響を与えるようにその構成を変更するか、または符号化する分布を変更することによって、自由エネルギーを最小化することができます。これらの変更はそれぞれ、行動（action）と知覚（perception）に対応し、生物学的システムに特徴的な環境との適応的な交換につながります。この扱いは、システムの状態と構造が、環境の暗黙的かつ確率的なモデルを符号化していると仮定しています。本稿では、脳によって内包されるモデルと、その自由エネルギーの最小化がそのダイナミクスと構造をどのように説明できるかについて考察します。

1 序論

概念的および数学的モデルを構築する私たちの能力は、周囲の世界を科学的に説明する上で中心的な役割を果たします。神経科学は、このモデル作成プロセスそのもののモデルを内包するという点で、他に類を見ません。知覚的にも科学的にも、世界に関する私たちの推論が、その推論を行うプロセスそのものに適用できるという事実は、実に注目に値します。現在、多くの人々は脳を、科学データの調査を司るのと同じ原理に従う推論機械であると考えています（MacKay, 1956; Neisser, 1967; Ballard et al., 1983; Mumford, 1992; Kawato et al., 1993; Rao and Ballard, 1998; Dayan et al., 1995; Friston, 2003; Körding and Wolpert, 2004; Kersten et al., 2004; Friston, 2005）。日常生活において、これらの規則は、私たちの感覚で世界をサンプリングすることで得られた情報に適用されます。過去数

年にわたり、私たちはこの見方をベイジアンフレームワークで追求し、脳がその感覚の原因を推論するために階層的または経験的ベイズを用いることを示唆してきました（Friston, 2005）。階層的な側面は、脳が自身の事前確率を学習し、暗黙的に感覚データを生成する内在的な因果構造を学習できるため、重要です。この脳機能モデルは、脳システムの幅広い解剖学的および生理学的側面を説明できます。例えば、皮質領域の階層的な配置、順方向および逆方向接続を用いた再帰的なアーキテクチャ、これらの接続における機能的非対称性などです（Angelucci et al., 2002a; Friston, 2003）。シナプス生理学の観点からは、連合性可塑性を予測し、動的モデルの場合にはスパイクタイミング依存性可塑性を予測します。電気生理学の観点からは、古典的および非古典的受容野効果、および誘発された皮質応答の長い潜時または内因性成分を説明します（Rao and Ballard, 1998; Friston, 2005）。これは、知覚学習に伴う予測誤差を符号化する応答の減衰を予測し、反復抑制、ミスマッチ陰性電位（MMN）、脳波における P300 など、多くの現象を説明します。精神物理学的な観点からは、プライミングやグローバル・プレシデンスなど、これらの生理学的現象の行動的相関を説明します（概要については Friston, 2005 を参照）。

知覚推論と学習の両方が自由エネルギーの最小化（Friston, 2003）または予測誤差の抑制（Rao and Ballard, 1998）に基づいていることは、比較的容易に示せます。自由エネルギーの概念は統計物理学に由来し、推論に内在する困難な積分問題をより簡単な最適化問題に変換するために、機械学習で広く用いられています。この最適化、すなわち自由エネルギー最小化は、原理的には比較的単純なニューロンのインフラストラクチャを用いて実装できます。

本稿の目的は、推論が自由エネルギー最小化の単なる創発的側面の一つであり、脳における自由エネルギー原理が知覚と行動の密接な関係を説明できることを示唆することです。さらに、自由エネルギー原理によってもたらされるプロセスは、世界の現在の状態に関する推論だけでなく、注意および関連メカニズムの特徴を示す文脈の動的な符号化も網羅します。

自由エネルギー原理は、システムがその自由エネルギーを減少させるように変化すると述べています。自由エネルギーの概念は、特に物理学と統計学において多くの文脈で生じます。熱力学では、自由エネルギーはシステムから取り出すことができる仕事量の尺度であり、工学用途で有用です。それはシステムのエネルギーとエントロピーの差です。自由エネルギーは統計学においても中心的な役割を果たし、統計熱力学から借用して、変分自由エネルギー最小化による近似推論（変分ベイズまたはアンサンブル学習としても知られる）は、最尤法と最大事後確率法を特殊なケースとして含みます。私たちが生物学的システムと世界との間の間の交換に適用するのは、この種の自由エネルギー、すなわち統計的確率分布の尺度です。これは、これらのシステムが周囲について暗黙的な推論を行うことを意味します。推論における自由エネルギーの以前の扱い（例：予測符号化）は、説明またはメカニズム記述として枠付けされてきました。本研究では、自由エネルギー最小化が生物学的システムにおいて必須であり、したがってより根本的な地位を持つことを示唆することによって、さらに一歩進もうと試みます。私たちはこれを、理論生物学と統計熱力学から引き出された一連のヒューリスティクスを提示することによって行います。

2　概要

本稿は 3 つのセクションで構成されています。最初のセクション（セクション 3〜7）では、選択主義的な視点から始め、自由エネルギー原理の神経生物学的および認知的な意味合いで締めくくることで、自由エネルギー原理の背後にある理論を説明します。第二のセクション（セクション 8〜10）では、階層的なニューロンアーキテクチャにおける自由エネルギー最小化の実装について述べ、感覚誘発反応の簡単なシミュレーションを提供します。これは、自由エネルギー原理に従って自己組織化する脳のようなシステムのいくつかの主要な振る舞いを例示します。主要な現象、すなわち、高次皮質領域からのトップダウン予測による予測誤差の抑制は、第三のセクションで検討されます。この最後のセクション（セクション 11）では、神経生物学的研究が自由エネルギー原理にどのように対処するために使用されているかの一例に焦点を当てます。この例では、ヒト被験者の機能的磁気共鳴画像法（fMRI）を用いて、予測可能および予測不可能な刺激に対する視覚誘発反応を調べます。

3　理論

このセクションでは、生物学的システム、特に脳のための変分自由エネルギー原理につながる一連のヒューリスティクスを展開します。私たちは、生物学的システムがどのように機能するかという難しい問いを、その行動に対する制約に関するより単純な問いへと変換する、進化的または選択主義的な考察から始めます。これらの制約は、システムの状態によって符号化されるアンサンブル密度という重要な概念へと私たちを導きます。この密度は、環境と交換しているあらゆるシステムのために自由エネルギーを構築するために用いられます。次に、この自由エネルギーを、システムの（すなわち脳の）状態を決定する量、そして決定的に、環境に対する行動に関して最小化することの意味合いを考察します。この最小化が、世界に関する知覚推論、知覚的文脈の符号化（すなわち注意）、環境の因果構造に関する知覚学習、そして最後に、その環境との原理に基づいた交換、またはサンプリングに自然につながることがわかるでしょう。

自由エネルギー原理（すなわち、脳はその自由エネルギーを最小化するように変化する）の下では、自由エネルギーは脳にとってのリアプノフ関数となります。リアプノフ関数は、システムの状態で時間の経過とともに減少するスカラー関数です。ニューラルネットワークの文献では、口語的にハーモニー関数とも呼ばれます（Prince and Smolensky, 1997）。時間依存偏微分方程式の文献には、関連するエネルギー汎関数の多くの例があります（例：Kloucek, 1998）。通常、システムの構造と行動が与えられたときに、リアプノフ関数を推論しようとします。以下では、その逆の問題に取り組みます。リアプノフ関数が与えられたとき、自由エネルギーを最小化するシステムはどのようなものになるでしょうか？

3.1　生物学的システムの性質

生物学的システムは、エネルギーとエントロピーを環境と交換するという意味で、熱力学的に開放的です。さらに、それらは平衡から遠く離れて機能し、散逸的であり、自己組織化する振る舞いを示します（Ashby, 1947; Nicolis and Prigogine,

1977; Haken, 1983; Kauffman, 1993)。しかし、生物学的システムは単に散逸的な自己組織化システムである以上のものです。それらは、変化する、または非定常的な環境と交渉し、かなりの期間存続することを可能にします。この存続は、そうでなければ物理的構造を変化させる相転移を回避することを意味します（興味深い例外は、発達経路における相転移です。例えば、変態する昆虫の場合)。生物学的システムの重要な側面は、環境に作用して、その内部での位置や関係を変化させ、極端な温度、圧力、その他の外部場の影響を排除することです。環境を選択的にサンプリングまたはナビゲートすることにより、物理的完全性を維持し、より長く存続できる範囲内で環境との交換を制限します。

図 1 に空想的な例を示します。ここでは、非生物的な自己組織化システムの典型例である雪の結晶を取り上げ、環境に作用できるように翼を与えています。通常の雪の結晶は落下し、相境界に遭遇し、そこで環境の温度によって溶けてしまいます。対照的に、高度を維持し、温度を調節できる雪の結晶は、質的に認識可能な形を保ちつつ、無限に存続します。通常の雪の結晶と適応的な雪の結晶との間の主要な違いは、環境との関係を変化させ、熱力学恒常性を維持する能力にあります。同様のメカニズムは、進化的な設定においても容易に想定できます。そこでは、相転移を回避するシステムが、できないシステムよりも選択されるでしょう（例：単細胞生物における化学走性や植物の光走性の選択と比較してください)。選択圧の観点から生物学的システムの性質を考察することで、生物学的システムがどのように出現するかという難しい問いを、存在するためにどのような行動を示さなければならないかという問いに置き換えることができます。言い換えれば、選択は生物学的システムがどのように生じるかを説明します。残された唯一の問題は、それらがどのような特性を持たなければならないかです。雪の結晶の例は、生物学的システムが相転移を排除するために環境に作用することを示唆しています。したがって、この種の環境との交換を保証する原理を定義すれば十分です。私たちは、自由エネルギー最小化がそのような原理の一つであることを知るでしょう。

3.2 アンサンブル密度

これらの議論を形式的に展開するために、環境、システム、およびそれらの相互作用を記述するいくつかの量を定義する必要があります。環境に作用する力または場をパラメータ化する量を ϑ とし、システムの物理的状態を記述する量を λ とします。これらの量については後で詳しく説明しますが、現時点では、これらが非常に高次元で時間変化する可能性があることに留意してください。

これらの 2 つの量を関連付けるために、アンサンブル密度と呼ぶ任意の関数 $q(\vartheta;\lambda)$ を使用します。これは、システムのパラメータによって指定または符号化される、環境のパラメータに関する任意の密度関数です。たとえば、λ は環境温度 ϑ のガウス分布の平均と分散である可能性があります。$q(\vartheta;\lambda)$ がアンサンブル密度と呼ばれる理由は、システムの定まった状態 λ が与えられたときに、特定の環境状態 ϑ が無限の環境アンサンブルから選択される確率密度と見なせるためです。ここで、$q(\vartheta;\lambda)$ は条件付き密度ではないことに注意してください。これは、λ が固定され既知であると扱われるためです（確率変数とは対照的に)。

アンサンブル密度を導入することで、システムの状態を環境と関連付け、システムを環境の確率モデルとして解釈できます。アンサンブル密度は、後述する自由エネルギーの定式化において中心的な役割を果たします。この定式化を説明

する前に、それぞれ環境がシステムに与える影響とシステムが環境に与える影響を記述する、他の 2 つの変数セットについて考察する必要があります。これらをそれぞれ $\tilde{y}$ と α とします。$\tilde{y}$ は、環境によって作用されるシステムの状態、たとえば感覚受容器の状態と考えることができます。これは、$\tilde{y}$ が感覚入力と見なせることを意味します。作用変数 α は、感覚サンプルの変化のために環境に作用するエフェクターによって加えられる力を表します。この依存性は、感覚サンプル $\tilde{y}(\alpha)$ を作用の汎関数とすることで表現します。

この依存性は、単純な場合があります。例えば、筋紡錘の伸張受容器の活動は、その紡錘を収縮させる筋力によって直接影響を受けます。他方、依存性がより複雑な場合もあります。例えば、眼の位置を制御する動眼神経系は、網膜のすべての光受容器の活動に影響を与えることができます。図 2 は、これらの変数とそれらがどのように関連しているかの概略図を示しています。これらの量が定まったところで、システムの自由エネルギーの表現を定式化できます。

4 自由エネルギー原理

自由エネルギーは、アンサンブル密度と現在の感覚入力のスカラー関数です。これは以下の 2 つの項から構成されます。

$$F = -\int q(\vartheta) \ln \frac{p(\tilde{y}, \vartheta)}{q(\vartheta)} d\vartheta = -\langle \ln p(\tilde{y}, \vartheta) \rangle_q + \langle \ln q(\vartheta) \rangle_q$$

最初の項は、アンサンブル密度のもとで期待されるシステムのエネルギーです。このエネルギーは、感覚入力とその原因 ϑ が同時に生起することについての驚き、または情報に過ぎません。第二の項は、アンサンブル密度の負のエントロピーです。自由エネルギーは 2 つの密度によって定義されることに注意してください。すなわち、アンサンブル密度 $q(\vartheta; \lambda)$ と、生成密度 $p(\tilde{y}, \vartheta)$ と呼ぶもので、これから感覚サンプルとその原因を生成することができます。生成密度は、尤度と事前密度 $p(\tilde{y}|\vartheta)p(\vartheta)$ に因数分解され、これにより生成モデルが特定されます。

これは、自由エネルギーがあらゆるシステムに対して生成モデルを、そしてそのモデルの原因またはパラメータに対するアンサンブル密度を誘導することを意味します。自由エネルギーを評価するためには、これらの密度の関数形式が必要です。脳が採用する可能性のある関数形式については、次のセクションで考察します。現時点では、これらの関数形式により、自由エネルギーがシステムの感覚入力と状態の関数 $F(\tilde{y}, \lambda)$ として定義できることだけを記しておきます。

自由エネルギー原理は、変化しうる量、すなわちシステムが所有するすべての量が、自由エネルギーを最小化するように変化すると述べています。これらの量は、システムパラメータ λ と作用パラメータ α です。この原理は、後述するように、相転移を排除する環境との適応的な交換を説明するのに十分です。私たちは、λ と α に関して自由エネルギーを最小化することの意味合いを考慮することによって、このことを示します。

システムのパラメータを自由エネルギーに関して最適化することで、アンサンブル密度が、感覚データが与えられた環境的原因の事後確率または条件付き密度となることは、比較的容易に示せます。これは、自由エネルギーの λ への依存性を示すために、式 (1) を並べ替えることで確認できます。

$$F = -\ln p(\tilde{y}) + D(q(\vartheta;\lambda)\|p(\vartheta|\tilde{y}))$$

第二項のみが λ の関数です。これは、カルバック・ライブラー交差エントロピーまたはダイバージェンス項であり、アンサンブル密度と原因の条件付き密度との間の差を測定します。この尺度は常に正であるため、自由エネルギーを最小化することは、アンサンブル密度を条件付き密度と同じにすることに相当します。言い換えれば、システムの状態で符号化されたアンサンブル密度は、その感覚入力の原因の事後確率の近似となります。これは、システムがその感覚サンプルの原因を暗黙的に推論または表現していることを意味します。明らかに、この近似はシステムの物理的構造とアンサンブル密度の暗黙的な形式、そしてそれが環境の因果構造とどれだけ密接に一致するかに依存します。

ここでも、選択主義的な議論を援用すると、自分が置かれている環境の外部の因果構造に内部構造を一致させるシステムは、より効果的に自由エネルギーを最小化できるでしょう。

4.1 行動：α の最適化

作用変数に関して自由エネルギーを最小化することによって、システムの設定を変更し、環境を移動または再サンプリングすることは、アンサンブル密度と一致する環境のサンプリングを強制します。これは、自由エネルギーが α にどのように依存するかを示す、式 (1) の 2 番目の再編成から見ることができます。

$$F = -\langle \ln p(\tilde{y}(\alpha)|\vartheta)\rangle_q + D(q(\vartheta)\|p(\vartheta))$$

この場合、最初の項のみが作用の関数です。この項を最小化することは、アンサンブル密度の下で期待される感覚入力の対数確率を最大化することに相当します。言い換えれば、システムはアンサンブル密度の下で最も可能性の高い感覚入力をサンプリングするように、自身の配置を変更します。しかし、前述のように、アンサンブル密度は感覚入力が与えられた原因の条件付き分布に近似します。この本質的な循環性により、システムは自身の期待を満たすことを余儀なくされます。つまり、システムは、遭遇することを期待する環境内の原因に選択的に自身を曝露します。

ただし、これらの期待は、システムが占有できる物理的状態のレパートリーに限定され、それがアンサンブル密度を特定します。したがって、自由エネルギーが低いシステムは、自身の物理的状態のレパートリーで符号化できる環境の一部しかサンプリングできません。自由エネルギーが低いため、推論された原因は実際の環境条件に近似します。これは、システムが自身の存在証明であるため、システムの物理的状態がこれらの環境力の下で持続可能でなければならないことを意味します。

要するに、自由エネルギーの低いシステムは、外部または内部環境の変化に適応的に反応し、環境との恒常的な交換を維持しているように見えるでしょう。

自由エネルギーを最小化できないシステムは、アンサンブル密度を表すための最適ではない構造、または環境をサンプリングするための不適切なエフェクターを持つことになります。これらのシステムは、自身の環境の特定の領域に自身を限定せず、最終的には相転移を経験するでしょう。

まとめると、自由エネルギー原理は、自由エネルギーを最小化しないシステムは環境変化に反応できず、「生物学的」という属性を持つことができないということに注目するだけで、非常に簡単に動機づけられます。したがって、自由エネルギーの最小化は、十分ではないにしても、必要な生物学的特性である可能性があります。生物学的システムが自由エネルギーを最小化するメカニズムは、選択圧に起因すると考えられます。これは、体細胞的（すなわち、生物の寿命）または進化的タイムスケールで作用します（Edelman, 1993）。脳における自由エネルギーの最小化に目を向ける前に、私たちは生物学的システムを記述する量を解きほぐし、それらのダイナミクスを神経科学におけるプロセスと関連付ける必要があります。

4.2 平均場近似

これまで、環境を記述する量をひとまとめにして扱ってきました。もちろん、これらの量はその数も種類も膨大です。それらの間の重要な違いは、変化する時間スケールです。この違いを利用して、パラメータを 3 つのセットに分割します。$\vartheta = \vartheta_u, \vartheta_\gamma, \vartheta_\theta$ はそれぞれ、速く、遅く、非常に遅く変化するものであり、アンサンブル密度を次のように因数分解します。

$$q(\vartheta) = \prod_i q(\vartheta_i; \lambda_i) = q(\vartheta_u; \lambda_u) q(\vartheta_\gamma; \lambda_\gamma) q(\vartheta_\theta; \lambda_\theta)$$

これはまた、システムパラメータを $\lambda = \lambda_u, \lambda_\gamma, \lambda_\theta$ に分割することを誘導します。これらはアンサンブル密度の時間変化する分割を符号化します。最初のセット λ_u は、急速に変化するシステム量です。これらは、神経活動や脳の電磁状態に対応し、ミリ秒のタイムスケールで変化します。それらが符号化する原因 ϑ_u は、急速に変化する環境状態、例えば構造的不安定性や他の生物によって引き起こされる環境の変化に対応します。

2 番目のセット λ_γ は、秒単位のタイムスケールでよりゆっくりと変化します。これらは、ニューロンにおける分子シグナル伝達の動力学に対応します。例えば、シナプス効率の短期的な変化や古典的な神経調節効果の根底にあるカルシウム依存性メカニズムなどです。環境における原因の同等な分割は、放射照度レベルや、その状態のより急速な変動の文脈を設定するゆっくりと変化する外部場の影響など、文脈的な性質のものである可能性があります。

最後に、λ_θ はゆっくりと変化するシステム量を表します。例えば、経験依存性可塑性におけるシナプス結合の長期的な変化や、神経発達的なタイムスケールで変化する軸索の展開などです。環境における同種の量は、因果的アーキテクチャにおける不変性です。これらは、物理法則や、世界との相互作用を形作る他の構造的規則性に対応する可能性があります。

式 (4) の因数分解は、統計物理学では平均場近似として知られています。明らかに、3 つの分割を用いた私たちの近似はやや恣意的ですが、神経系におけるそれぞれの最適化の機能的相関を整理するのに役立ちます。植物のような他のシステムには、他のタイムスケールが必要になるでしょう。平均場近似は、特定のスキームを考慮する際に、自由エネルギーの最小化を大幅に微調整します。これらのスキームは通常、変分法を用います。変分法は、ファインマン (1972) によって経路積分定式化を用いた量子力学の文脈で導入されました。それらは機械

学習コミュニティで広く採用されています（例：Hinton and von Cramp, 1993; MacKay, 1995）。期待値最大化や制限付き最尤法といった確立された統計的手法（Dempster et al., 1977; Harville, 1977）は、自由エネルギーの観点から定式化できます（Neal and Hinton, 1998; Friston et al., in press）。

5　変分モードの最適化

次に、平均場近似を用いて、知覚の根底にあるシステムパラメータの最適化を、さらに詳しく見ていきます。この近似の下では変分法が主流であるため、式 (1) の自由エネルギーは変分自由エネルギーとも呼ばれ、λ_i は変分パラメータと呼ばれます。平均場分解は、平均場近似が、あるパーティションにおけるランダムなゆらぎが、別のパーティションにおけるゆらぎに与える影響をカバーできないことを意味します。しかし、これらの影響は平均場効果（すなわち、ランダムなゆらぎの平均や分散を介して）によってモデル化されるため、これは深刻な制限ではありません。この近似は、速いタイムスケールでのランダムなゆらぎが遅いタイムスケールで直接的な影響を及ぼす可能性が低いという点で、現在のフレームワークでは特に動機づけやすいです。

変分法を用いると、（Friston et al., in press を参照）上記の平均場近似の下で、アンサンブル密度が以下の形式を持つことを簡単に示すことができます。

$$q(\vartheta_i) \propto \exp(I(\vartheta_i))$$
$$I(\vartheta_i) = \langle \ln p(\tilde{y}, \vartheta) \rangle_{q_{\setminus i}}$$

ここで、$I(\vartheta_i)$ は、ϑ_i と、他のパーティションのアンサンブル密度 $q_{\setminus i}$ の下で期待されるデータの単純な対数確率です。これを変分エネルギーと呼びます。式 (5) から、アンサンブル密度のモードが変分エネルギーを最大化することが明らかです。モードは重要な変分パラメータです。例えば、$q(\vartheta_i)$ がガウス分布であると仮定すると、それはモードと共分散をそれぞれ符号化する 2 つの変分パラメータ $\lambda_i = \mu_i, \Sigma_i$ によってパラメータ化されます。これはラプラス近似として知られており、後で用いられます。以下では、μ_i を最適化することによって自由エネルギーを最小化することに焦点を当てます。各パーティションについて、アンサンブル密度の高次モーメントを記述する他の変分パラメータが存在する可能性があることに注意してください。幸いなことに、ラプラス近似の下では、私たちが必要とする唯一の他の変分パラメータは共分散です。これは単純な形式を持ち、モードの解析関数であるため、明示的に表現する必要はありません（Friston et al., in press および付録 A を参照）。

次に、変分モード μ_i の最適化と、この最適化が伴う神経生物学的および認知プロセスを見ていきます。

5.1　知覚推論：μ_u の最適化

ニューロン状態 μ_u に関して自由エネルギーを最小化することは、$I(\vartheta_u)$ を最大化することを意味します。

$$\mu_u = \max I(\vartheta_u)$$

$$I(\vartheta_u) = \langle \ln p(\tilde{y}|\vartheta) + \ln p(\vartheta) \rangle_{q_\gamma q_\theta} = \langle \ln p(\vartheta|\tilde{y}) \rangle_{q_\gamma q_\theta} + \ln p(\tilde{y})$$

これは、自由エネルギー原理が機能する際、状態の変分モード（すなわち、ニューロン活動）が、よりゆっくりと変化する原因のアンサンブル密度のもとで期待される対数事後確率を最大化するように変化することを意味します。これは、真の事後確率を知ることなく、確率的生成モデルを特定する期待される対数尤度と事前確率を最大化することによって達成されます（式 (6) の 2 行目）。

前述のように、この最適化には生成モデルの関数形式が必要です。次のセクションでは、脳の構造と両立する階層形式について考察します。今のところは、自由エネルギー原理が、脳の状態が感覚入力の原因となっている環境の最も可能性の高い状態を符号化するようになる、ということを理解していれば十分です。

5.2 一般化座標

状態は時間変化する量であるため、そのアンサンブル密度が何をカバーするのかを考慮することが重要です。これは、ある時点での状態だけでなく、その高次の運動もカバーすることができます。言い換えれば、環境の特定の状態とその脳内での確率的符号化は、一般化座標における状態の軌跡を表現することによってダイナミクスを具現化することができます。一般化座標は物理学における一般的な手法であり、通常、位置と運動量をカバーします。現在の文脈では、一般化された状態には、現在の状態とその一般化された運動 $\vartheta_u = u, u', u'', \ldots$ が含まれ、対応する変分モードは $\mu_u, \mu'_u, \mu''_u, \ldots$ となります。式 (6) における極値化は、高次から低次への運動を平均場項を介して結合しながら、迅速な勾配降下によって達成できることを示すのは比較的簡単です（Friston, in preparation）。

$$\begin{aligned}
\dot{\mu}_u &= \kappa \partial I(\vartheta_u)/\partial u + \mu'_u \\
\dot{\mu}'_u &= \kappa \partial I(\vartheta_u)/\partial u' + \mu''_u \\
\dot{\mu}''_u &= \kappa \partial I(\vartheta_u)/\partial u'' + \mu'''_u \\
\dot{\mu}'''_u &= \cdots
\end{aligned}$$

ここで、$\dot{\mu}_u$ は μ_u の変化率を意味し、κ は適切なレート定数です。次のセクションのシミュレーションでは、この降下スキームを使用します。これは、比較的単純なニューラルネットワークを使用して実装できます。条件付きモードが $I(\vartheta_u)$ の最大値を見つけた場合、その勾配はゼロになり、モードの運動は運動のモードになることに注意してください。つまり、$\dot{\mu}_u = \mu'_u$ です。しかし、特別な制約がない限り、一般化座標ではこれらの量が異なることは十分に可能です。知覚のレベルでは、運動残効のような精神物理学的現象は、脳が一般化座標を使用することを示唆しています。例えば、動いている列車から景色をしばらく見た後、立ち止まると、世界が動いているように知覚されるものの、位置は変化しないといった現象です。視覚オブジェクトがその動きに合わせて位置を変えるという印象は、私たちの脳が世界の因果構造について学習したことです。また、これは一時的に学習され直すことも可能です（例：知覚残効）。

次に、これらの因果的規則性がどのように学習されるかについて考察します。

5.3 知覚の文脈と注意：μ_γ の最適化

中間的なタイムスケールで変化する原因を ϑ_γ とし、これを文脈的と呼ぶなら、μ_γ を最適化することは、状態の速いダイナミクスがどのように進化するかの確率的偶発性を符号化することに相当します。この最適化は前述と同様に進めることができますが、ここでは文脈が十分にゆっくりと変化するため、$\mu'_\gamma = 0$ と近似できると仮定します。これにより、単純な勾配上昇が得られます。

$$\dot{\mu}_\gamma = \kappa \partial I(\vartheta_\gamma)/\partial \vartheta_\gamma$$
$$I(\vartheta_\gamma) = \langle \ln p(\tilde{y}, \vartheta) \rangle_{q_u q_\theta}$$

この期待値は、状態の一般化座標、そして暗黙的に、状態軌跡が進化する長い時間期間にわたるものであることに注意してください。後述するように、文脈を符号化する条件付きモード μ_γ は、脳内のニューロン間の側方または水平な相互作用の強さに対応する可能性があります。これらの側方相互作用は、期待される状態に対するトップダウンとボトムアップの影響の相対的な効果を制御し、したがって、知覚推論を行う上での経験的事前確率と感覚情報との間のバランスを制御します。これは、注意がこのような文脈的パラメータの最適化として捉えられる可能性を示唆しています。

式 (8) では、μ_γ のダイナミクスが知覚状態のアンサンブル密度のもとでの期待によって決定されることが重要です。これは、システムが、状態とその履歴に敏感な方法で、確率的偶発性の内部表現を調整できることを意味します。心理学におけるこの単純な例として、ポズナー・パラダイムが挙げられます。ここでは、知覚状態、すなわち定位キューが、ターゲットキューが提示される視覚空間の特定の領域に視覚的注意を向けます。現在の定式化では、これは感覚器のキューされた部分へと知覚推論を偏らせる、文脈を符号化する変分パラメータの状態依存的変化に対応するでしょう（これについては今後の論文でモデル化します）。

ここでの重要な点は、平均場近似が、急速に変化する知覚状態とよりゆっくりと変化する文脈に関する推論が、平均場効果（すなわち、式 (6) および (8) の期待値）を通じて相互に影響し合うことを可能にする点です。これは、状態と文脈に関する共同分布をアンサンブル密度で明示的に表現することなく進めることができます（Rao, 2005 参照）。変分パラメータ間のもう一つの重要な相互作用は、不確実性の符号化に関連しています。ラプラス近似の下では、これは条件付き共分散によって符号化されます。決定的に重要なのは、あるアンサンブルの条件付き共分散が、他のアンサンブルの条件付きモードの関数であることです（付録 A の式 (A.2) を参照）。現在の文脈では、知覚推論に対する文脈の影響は、不確実性の符号化という観点から捉えることができます。これのニューロン実装については、次のセクションで考察します。

5.4 知覚学習：μ_θ の最適化

ϑ_θ を符号化する変分モードを最適化することは、環境の因果的アーキテクチャにおける構造的規則性を推論し、学習することに相当します。前述と同様に、この学習は、一般化された状態と文脈を符号化するアンサンブル密度のもとでの期待を表す $I(\vartheta_\theta)$ の勾配上昇として実装できます。

$$\dot{\mu}_\theta = \kappa \partial I(\vartheta)/\partial \vartheta_\theta$$

$$I(\vartheta_\theta) = \langle \ln p(\tilde{y}, \vartheta) \rangle_{q_u q_\gamma}$$

脳内では、この降下は、シナプス前予測とシナプス後予測誤差の関数である結合の変化として定式化できます（Friston, 2003, 2005 参照）。結果として得られる学習ルールは、単純な連合性可塑性に、または動的モデルにおいてはスパイクタイミング依存性可塑性に類似した可塑性に適合します。脳内の結合強度に対応する変分パラメータを最適化することが、環境内の因果構造を符号化するという意味において、この自由エネルギー最小化のインスタンスは学習に相当します。

脳の結合における暗黙的な変化は、知覚と注意の根底にある自由エネルギーのダイナミクスに影響を与える、環境との過去の相互作用の記憶を付与します。これは、式 (6) および (8) における平均場効果を介して行われます。簡単に言えば、環境入力への継続的な曝露は、脳の内部構造にそれらの入力の因果構造を再現させます。これにより、効率的な知覚推論が可能になります。この定式化は、システムが因果状態と文脈間の関連性や偶発性を記憶することを可能にする、知覚学習とカテゴリ化について明快な説明を提供します。これらのアイデアをエピソード記憶へと拡張することは、未解決の課題として残っています。

6　モデルの最適化

これまで、特定のシステムとその暗黙的な生成モデルが与えられた上での、変分パラメータの定量的最適化のみを考察してきました。しかし、全く同じ自由エネルギー原理を、モデル自体を最適化するためにも適用できます。異なるモデルは、システムの集団から、または時間の経過に伴う単一システムの質的変化から生じる可能性があります。

ここでのモデルは、同じ変分パラメータのセットで列挙できる特定の構成に対応します。例えば、システムの一部を取り除くことや、別の結合セットを追加することは、モデルと変分パラメータを質的またはカテゴリ的に変化させます。

モデルの最適化は、モデル自体の周辺尤度を最大化することを含みます。統計学や機械学習において、これはベイズモデル選択と同等であり、自由エネルギーは特定のモデル m_i の周辺尤度 $p(\tilde{y}|m_i)$、すなわちエビデンスを近似するために使用できます。この近似は、式 (2) を用いることで簡単に説明できます。もしシステムが自由エネルギーを最小化し、ダイバージェンス項がほぼゼロであるならば、自由エネルギーはそのモデルの負の対数エビデンスに近づきます。したがって、最も自由エネルギーが小さいモデルが、最も高い周辺尤度を持ちます。

この点に関する進化的視点では、対数エビデンスを自由エネルギーの下限と見なします。自由エネルギーは、システムの環境との交換 $\tilde{y}(\alpha)$ に対して定義され、システムのパラメータ λ には依存しません。適応システムは、その物理的完全性を確保する範囲内でこの交換を維持します。これに必要なのは、これらの範囲内で対数エビデンスを凹関数にし、その自由エネルギーを最小化するプロセスをもたらす適切なモデルを選択することだけです（図 2 参照）。最も低い自由エネルギーを持つモデルを選択することは、環境ニッチを最もよくモデル化でき、したがってその中に留まることができるモデルを選択することを意味します。

この階層的な選択は、特定のモデルのパラメータを最適化すること（自由エネルギーを最小化するため）と、モデルそのものを最適化すること（最小化された自由エネルギーを使用すること）の相互作用に基づいていることに注意してください。両方のレベルでの最適化は、自由エネルギー原理によって規定されています。遺伝的アルゴリズムの理論では、これは階層的共進化と呼ばれます（例：Maniadakis and Trahanias, 2006）。同様の関係は、ベイズ推論にも見られます。そこでは、モデル選択は、各モデルのパラメータを最適化して自由エネルギーを最小化することによって提供される、モデルエビデンスへの自由エネルギー近似に基づいています。要するに、自由エネルギーは、進化的設定における適応度、およびモデル選択における周辺尤度の有用な代替となる可能性があります。私たちがモデル選択を導入するのは、それが価値学習（図 3）と関連しているからです。

6.1 価値学習：m_i の最適化

ここでの価値学習は、システムが価値のある、または適応的な応答を学習する能力を指します。これは、心理学の文献では再強化学習（re-enforcement learning）または情動学習（emotional learning）と呼ばれ、工学および神経科学の文献では動的計画法（dynamic programming）（例：時間差分モデル）と密接に関連しています（例：Montague et al., 1995; Suri and Schultz, 2001）。価値学習の初期の定式化（Friston et al., 1994）では、私たちは生得的価値（innate value）と獲得的価値（acquired value）の区別を導入しました。生得的価値とは、適合度を高める遺伝的またはエピジェネティックに指定された応答を引き起こす刺激や感覚入力の属性です。獲得的価値とは、最終的に生得的価値を持つ刺激や手がかりを明らかにする行動を誘発するようになる刺激の属性です。したがって、獲得的価値は神経発達および環境への曝露の間に学習されます。

自由エネルギー原理は、獲得的価値や再強化の概念を援用することなく、適応的な行動を説明します。生物の観点から見ると、それは単に感覚入力が自身の期待に合致するように環境をサンプリングしているに過ぎません。その視点から見れば、環境は安定して順応的な場所です。しかし、このシステムを観察する者にとっては、環境の変化に適応的に反応し、不利な条件を避けているように見えるでしょう。言い換えれば、内部環境の恒常性を確保するために、特定の刺激と応答の結びつきが選択的に再強化されているかのように見えるでしょう。しかし、この強化は、自由エネルギー原理の下での行動と知覚というより大きな文脈において自発的に出現します。簡単な例として、「暗闇を好む」昆虫を想像してください。世界は暗いものだと期待するように進化した昆虫を想像してみてください。その昆虫は、常に暗い環境をサンプリングできるように、影の中へ移動するでしょう。観察者の視点から見ると、この適応的な行動は、「影」の価値によって強化された光回避行動と [誤って] 解釈されるかもしれません。

上記の議論は、生物学的システムが、その構造に内在するモデルによって生成された期待を満たすために環境をサンプリングすることを示唆しています。そのモデルの尤度部分は、環境への曝露によって学習されます。しかし、その事前確率は遺伝によって受け継がれる可能性があります。生得的価値に対応し、モデル m_i そのものの一部であるのは、これらの事前確率です。価値学習はしばしば、期待される報酬や価値を最大化するという観点から捉えられます。しかし、自由エネルギーの定式化ではこれは必要ありません。必要なのは、生物が自身の期待を最大化することだけです。選択は、生得的に価値のある事前確率に対する選択

圧を通じて、これらの期待が価値あるものであることを保証するでしょう。擬人化して言えば、私たちは報酬を最大化するために世界と相互作用するのではなく、単に世界が私たちが考えるように振る舞うことを確実にするために相互作用するのかもしれません。良いモデルと、そもそも良い世界のモデルを持つ表現型だけが生き残るでしょう。悪いモデルを持つか、そもそも悪い世界をモデル化する者は絶滅するでしょう。

まとめると、生物の生涯において、その表現型に内在するモデルが与えられたとき、そのパラメータは自由エネルギーを最小化します。より高次のレベルでは、モデル自体が選択され、集団がモデル空間を探索し、最適なモデルを見つけることを可能にします。この探索は、生物が機能できる環境ニッチに関する事前確率を含む、主要なモデルコンポーネントの遺伝性に依存します。

このセクションでは、生物の状態と構造の進化のための自由エネルギー原理を展開し、階層的選択を通じて集団レベルでの自由エネルギーの最小化についても触れました。自由エネルギーの最小化は、感覚入力の原因に関するアンサンブル密度をパラメータ化する生物の構成を最適化すること、および体細胞的または進化的時間においてモデル自体を最適化することに対応します。異なるタイムスケールで変化する量をカバーするためのアンサンブル密度の因数分解は、知覚推論、注意、学習にうまく対応するプロセスのオントロジーを提供します。もちろん、私たちはこれらの問題にやや表面的に触れたに過ぎません。それぞれが十分に扱われるべきです。

本稿では、知覚推論に焦点を当てます。次のセクションでは、脳がどのように自由エネルギー原理を実体化するのか、特にその構造によって内包される尤度モデルに焦点を当てて考察します。

7　脳における生成モデル

このセクションでは、前セクションの比較的抽象的な原理が脳にどのように適用され得るかを見ていきます。私たちはすでに、生物学的構造がそれが浸漬している環境のモデルを符号化するという考え方を導入しました。ここでは、脳の構造が示唆するこれらのモデルの形式と、誘発反応や連合性可塑性が自由エネルギーの最小化として自然にどのように生じるかを理解しようとします。

現在の定式化では、脳を記述するあらゆる属性や量は、環境の原因に関するアンサンブル密度をパラメータ化します。この密度の自由エネルギーを評価するには、アンサンブル密度と生成密度の関数形式を特定する必要があります。私たちは、アンサンブル密度に対してはガウス形式（すなわちラプラス近似）を仮定します。これは、そのモードまたは期待値と共分散によってパラメータ化されます。生成密度は、その尤度と事前確率によって特定されます。これらが組み合わさって生成モデルを構成します。このモデルが適切に特定されていれば、自由エネルギー原理を用いて、脳が異なる文脈でどのように振る舞うかを予測できるはずです。

これまでのいくつかの論文（例：Friston and Price, 2001; Friston, 2005）で、脳が採用する可能性のある階層的生成モデルの形式について記述してきました。このセクションでは、その主要な点を簡潔に再確認します。

7.1 脳における階層的動的モデル

脳の重要な構造原理は、その階層的組織です（Zeki and Shipp, 1988; Felleman and Van Essen, 1991; Mesulam, 1998; Hochstein and Ahissar, 2002）。この組織は視覚系で最も徹底的に研究されており、皮質領域が階層を形成すると見なせます。下位の領域は一次感覚入力に近く、上位の領域は多感覚的または連合的な役割を担っています。

階層の概念は、順方向接続（forward connections）と逆方向接続（backward connections）の区別に基づいています（Rockland and Pandya, 1979; Murphy and Sillito, 1987; Felleman and Van Essen, 1991; Sherman and Guillery, 1998; Angelucci et al., 2002a）。順方向接続と逆方向接続の区別は、脳における外因性接続の主な源泉および起源となる皮質層の特異性に基づいています。順方向接続は主に顆粒上層の表層錐体細胞から発生し、上位皮質領域の第四層または顆粒層の有棘星状細胞に終止します（Felleman and Van Essen, 1991; DeFelipe et al., 2002）。逆に、逆方向接続は主に顆粒下層の深層錐体細胞から発生し、下位皮質領域の顆粒下層および顆粒上層の細胞を標的とします。内在性接続は層内および層間両方に存在し、数ミリメートル離れたニューロン間の側方相互作用を仲介します。

外因性の順方向および逆方向接続の収束と発散により、上位領域の受容野は一般的に下位領域よりも大きくなります（Zeki and Shipp, 1988）。順方向接続と逆方向接続の間には重要な機能的区別があり、これにより逆方向接続はニューロン応答に対する効果がより変調的または非線形的になります（例：Sherman and Guillery, 1998）。これは、逆方向接続によって標的とされる顆粒上層における電圧感受性および非線形 NMDA 受容体の配置と一致しています。通常、逆方向接続のシナプスダイナミクスはより遅い時定数を持っています。このことから、順方向接続は駆動的であり、上位レベルで義務的な応答を引き起こす一方、逆方向接続は駆動的および変調的な両方の効果を持ち、より大きな空間的および時間的スケールで機能するという概念が生まれました。

脳の階層的構造は、感覚入力の階層的モデルを示唆しています。例えば：

$$
\begin{aligned}
y &= g(x^{(1)}, v^{(1)}) + z^{(1)} \\
\dot{x}^{(1)} &= f(x^{(1)}, v^{(1)}) + w^{(1)} \\
&\vdots \\
v^{(i-1)} &= g(x^{(i)}, v^{(i)}) + z^{(i)} \\
\dot{x}^{(i)} &= f(x^{(i)}, v^{(i)}) + w^{(i)} \\
&\vdots
\end{aligned}
$$

このモデルでは、感覚状態 y は、状態の非線形関数 $g(x^{(1)}, v^{(1)})$ にランダムな効果 $z^{(1)}$ が加わることによって引き起こされます。動的状態 $x^{(1)}$ は記憶を持ち、非線形関数 $f(x^{(1)}, v^{(1)})$ によって規定される運動方程式に従って進化します。これらのダイナミクスは、ランダムなゆらぎ $w^{(1)}$ と、全く同じ方法で生成される上位レベルからの摂動の影響を受けます。言い換えれば、いかなるレベルへの入力も、その上位レベルの出力です。これは、因果的状態 $v^{(i)}$ が階層レベルを結

びつけ、動的状態 $x^{(i)}$ が各レベルに内在するダイナミクスを生成することを意味します。ランダムなゆらぎは、ハイパーパラメータ $\vartheta_\gamma^{(i)}$ によって符号化された共分散を持つガウス分布であると仮定できます。各レベルの関数は $\vartheta_\theta^{(i)}$ によってパラメータ化されます。この形式の階層的動的モデルは極めて一般的であり、統計学や機械学習で見られるほとんどのモデルを特殊なケースとして包含します。

このモデルは、運動の一般化座標における生成密度の関数形式を特定し（付録 B 参照）、一般化された状態 $\vartheta_u^{(i)} = \tilde{x}^{(i)}, \tilde{v}^{(i)}$ に関するアンサンブル密度を誘導します。もし、ニューロン活動がこれらの状態の変分モード $\tilde{\mu}_u^{(i)} = \tilde{\mu}_v^{(i)}, \tilde{\mu}_x^{(i)}$ であり、モデルパラメータ $\vartheta_\gamma^{(i)}$ および $\vartheta_\theta^{(i)}$ の変分モードがシナプス効率または結合強度に対応すると仮定すると、式 (5) を用いてこれらのモードの関数として変分エネルギーを記述できます。ここで $y = \mu_v^{(0)}$ とします。

$$I(\tilde{\mu}_u) = -\frac{1}{2}\Sigma_i \tilde{\epsilon}^{(i)T} \Pi^{(i)} \tilde{\epsilon}^{(i)}$$

$$\tilde{\epsilon}^{(i)} = \begin{bmatrix} \tilde{\epsilon}_v^{(i)} \\ \tilde{\epsilon}_x^i \end{bmatrix} = \begin{bmatrix} \tilde{\mu}_v^{(i-1)} - \tilde{g}(\tilde{\mu}_u^{(i)}, \mu_\theta^{(i)}) \\ \tilde{\mu}_x'^{(i)} - \tilde{f}(\tilde{\mu}_u^{(i)}, \mu_\theta^{(i)}) \end{bmatrix}$$

$$\Pi(\mu_\gamma^{(i)}) = \begin{bmatrix} \Pi_z^{(i)} & \\ & \Pi_w^{(i)} \end{bmatrix}$$

ここで、$\tilde{\epsilon}^{(i)}$ は i 番目のレベルにおける状態の一般化予測誤差です。因果的状態と動的状態の運動の一般化予測は、それぞれ $\tilde{g}^{(i)}$ と $\tilde{f}^{(i)}$ です（付録 B 参照）。ここで、$\tilde{\mu}'^{(i)} = \mu_x'^{(i)}, \mu_x''^{(i)}, \mu_x'''^{(i)}, \ldots$ は $\tilde{\mu}_x^{(i)}$ の運動を表します。$\Pi(\mu_\gamma^{(i)})$ は、ランダムなゆらぎの精度であり、その振幅と滑らかさを制御します。簡略化のため、パラメータの条件付き共分散に依存する項は省略しました。これは期待値最大化で用いられるのと同じ近似です（Dempster et al., 1977）。

7.2 知覚推論のダイナミクスとアーキテクチャ

前述のように、知覚または知覚推論に暗黙的に含まれる状態をカバーするアンサンブル密度の最適化に焦点を当てます。式 (7) から、自由エネルギー原理の下でのニューロン活動のダイナミクスを記述する表現が得られます。

$$\dot{\tilde{\mu}}_u^{(i)} = h(\tilde{\epsilon}^{(i)}, \tilde{\epsilon}^{(i+1)})$$

$$= \tilde{\mu}_u'^{(i)} - \kappa \frac{\partial \tilde{\epsilon}^{(i)T}}{\partial \tilde{\mu}_u^{(i)}} \Pi^{(i)} \tilde{\epsilon}^{(i)} - \kappa \frac{\partial \tilde{\epsilon}^{(i+1)T}}{\partial \tilde{\mu}_u^{(i)}} \Pi^{(i+1)} \tilde{\epsilon}^{(i+1)}$$

これらのダイナミクスは、脳が感覚入力に曝露されたときにニューロン状態がどのように自己組織化するかを記述します。式 (12) の形式は非常に示唆的です。それは主に予測誤差の関数であり、すなわち、どのレベルにおいても世界の状態の期待値と、上位レベルの期待状態に基づいて予測されるものとの間の不一致です。

決定的に重要なのは、推論が下位レベルの予測誤差 $\tilde{\epsilon}^{(i)}$ と、問題のレベル $\tilde{\epsilon}^{(i+1)}$ の予測誤差のみを必要とすることです。これにより、条件付き期待値 $\tilde{\mu}_u^{(i)}$ は、予測誤差を説明するために逆方向接続によって伝えられるより良い予測を提供する

ように駆動されます。これこそが、自由エネルギーまたは予測誤差を抑制するために自己組織化する再帰的ダイナミクス、すなわち認識ダイナミクスの本質です。

決定的に、期待される状態の運動はボトムアップ予測誤差の線形関数です。これは生理学的に観察されることと全く同じです。つまり、ボトムアップの駆動入力は、他のボトムアップ入力に依存しない義務的な応答を上位レベルで引き起こします。実際、式 (12) における順方向接続は単純な形式をしています。

$$\frac{\partial \tilde{\epsilon}^{(i)T}}{\partial \tilde{\mu}_u^{(i)}} \Pi^{(i)} - \begin{bmatrix} -I \otimes g_v^{(i)} & -I \otimes g_x^{(i)} \\ -I \otimes f_v^{(i)} & D - (I \otimes f_x^{(i)}) \end{bmatrix}$$

これは、導関数 $g_x = \partial g / \partial x$（他の導関数も同様）のブロック対角の繰り返しから構成されます。D は、一般化された運動の内部の一貫性を保証する、その第一対角に単位行列を持つブロック行列です。接続は $\mu_\gamma^{(i)}$ によって符号化された精度によって変調されます。各レベル内の側方相互作用は、さらに単純な形式をしています。

$$\frac{\partial \tilde{\epsilon}^{(i+1)T}}{\partial \tilde{\mu}_u^{(i)}} \Pi^{(i+1)} = \begin{bmatrix} \Pi_v^{(i+1)} & 0 \\ 0 & 0 \end{bmatrix}$$

そして、そのレベルにおける原因の精度に還元されます。これらの相互作用の生物学的基盤については後述します。

式 (12) の形式により、予測誤差の源を表層錐体細胞に帰属させることができ、これらの細胞が予測誤差を符号化していると仮定できます。これは、階層のあるレベルから次のレベルへと順方向に伝達される唯一の量が予測誤差であり、表層錐体細胞が脳における順方向求心性線維の源であるためです。これは、これらの細胞が非侵襲的に測定できる脳波（EEG）信号の発生に主に責任があるため有用です。

予測誤差自体は、逆方向接続によって伝えられる予測と、当該レベルに内在するダイナミクスによって形成されます。これらの影響は、$\tilde{g}^{(i)}$ および $\tilde{f}^{(i)}$ に暗黙的に含まれる非線形性を具現化します。式 (11) を参照してください。これもまた、逆方向接続の非線形または変調的役割と完全に一致しており、この文脈では、推論された状態間の相互作用をモデル化して下位レベルの推論を予測します。暗黙的なニューロンアーキテクチャの模式図については図 4 を参照してください。

要するに、条件付きモードのダイナミクスは 3 つの項によって駆動されます。

1. 一般化座標を連結し、モードの運動が運動のモードに近似することを保証する項。 これにより、因果的ダイナミクスが内部的に一貫して表現されます。2. 下位レベルからの予測誤差に依存するボトムアップ効果の項。 これは尤度項と考えることができます。3. 現在のレベルでの予測誤差によって媒介される項。 これは経験的事前確率に対応し、トップダウン予測を使用して構築されます。

階層的モデルの重要な側面は、それらが自身の経験的事前確率を構築できることです。統計学の文献では、これらのモデルはパラメトリック経験ベイズモデル（Efron and Morris, 1973）として知られており、各レベルにおけるランダムなゆらぎの条件付き独立性に依拠しています（Kass and Steffey, 1989）。要約する

と、脳のどのレベルにおける知覚推論のダイナミクスも、上位レベルからのトップダウン事前確率によって調整されます。これはすべてのレベルで再現され、再帰的相互作用を通じて自己組織化を可能にし、階層全体で予測誤差を抑制することで自由エネルギーを最小化します。このようにして、上位レベルは下位レベルにガイダンスを提供し、複数レベルの記述において感覚入力の推論された原因の内部一貫性を確保します。

8 知覚的注意と学習

上記のダイナミクスは、感覚入力の最も可能性の高い原因を記述する、条件付きまたは変分モードの最適化を表しています。これが知覚推論であり、式 (10) で記述された階層的生成モデルのベイズ的逆問題解決に対応します。この簡略化されたスキームでは、条件付き共分散が無視されており、自由エネルギーの最小化は階層的予測誤差の抑制と同等です。全く同じ扱いが、条件付きモード μ_γ および μ_θ を符号化する外因性および内在性結合の変化にも適用できます。前述のように、これらのモードまたはシナプス効率の変化は、予測誤差の比較的単純な関数であり、連合性可塑性として認識できる形式につながります。例：これらの導出は、静的システムに対しては Friston (2005) で提供されています。

文脈変数は、知覚推論を調整する役割という点で興味深いものです。式 (12) は、下位レベルからの予測誤差と現在のレベルからの予測誤差の影響が、μ_γ の関数である精度行列 $\Pi(\mu_\gamma^{(i)})$ および $\Pi(\mu_\gamma^{(i+1)})$ によってスケーリングされることを示しています。これは、ボトムアップの尤度項とトップダウンの事前確率の相対的な影響が、μ_c によって符号化された変調的影響によって制御されることを意味します。この求心性線維の選択的変調は、注意のために援用されてきたゲイン制御メカニズムと全く同じです（例：Treue and Maunsell, 1996; Martinez-Trujillo and Treue, 2004）。各皮質レベルに内在する側方相互作用によってこのゲインが制御されるニューロンアーキテクチャを定式化することは、比較的簡単です（図 4 参照）。

前述のセクションで述べたように、μ_γ の変化は、状態の速いダイナミクスと、尤度モデルを媒介する外因性結合におけるゆっくりとした連合性変化との中間的なタイムスケールで発生するとされています。μ_γ は、古典的な神経調節性入力や他のゆっくりとしたシナプスダイナミクス（例：過分極後電位や分子シグナル伝達）に依存する、側方または内在性結合におけるシナプス効率の短期的な変化を記述すると考えることができます。これらの中間的ダイナミクスの生理学的側面は、脳における注意メカニズムにとって興味深い基質を提供し（Schroeder et al., 2001 のレビューを参照）、Yu and Dayan (2005) のアイデアと無関係ではありません。これらの著者らは、アセチルコリン（上行性変調性神経伝達物質）が期待される不確実性を媒介する役割を仮定しています。神経調節性神経伝達物質は、特徴的に、順方向および逆方向の外因性結合によって用いられるグルタミン酸性神経伝達よりも、シナプス効果の点ではるかに遅い時定数を持っています。

結論として、かなり一般的な階層的で動的な環境入力モデルが、自由エネルギーとその最小化を特定するためにニューロン量にどのように転写され得るかを見てきました。この最小化は、いくつかの簡略化された仮定の下で、皮質階層のすべてのレベルにおける予測誤差の抑制に対応します。この抑制は、ボトムアップ（尤度）の影響とトップダウン（事前確率）の影響との間のバランスに基づいて

おり、そのバランスは不確実性の表現によって取られています。そして、これらの表現は、古典的な神経調節効果や、予測誤差の全体的なレベルによって駆動される遅いシナプス後細胞プロセスによって媒介される可能性があります。全体として、これにより、文脈に敏感で自由エネルギー原理に適合する階層的感覚入力モデルのベイズ的逆問題解決が可能になります。次に、簡単なシミュレーションを用いて、脳内で見られると予想されるダイナミクスと行動の種類を説明します。

9　シミュレーション

9.1　生成モデルと認識モデル

ここでは、刺激が提示されたときの自己組織化されたダイナミクスの主要な特徴を示すために、2 層ニューロン階層の非常に単純なシミュレーションについて記述します。システムを図 5 に示します。左側は感覚入力を生成するために使用されるシステムであり、右側は、この生成を反転させる、すなわち、根底にある原因を認識または開示するために使用されるニューロンアーキテクチャです。

生成システムは、単一の入力（ガウスのバンプ関数）を使用し、これが相互に接続された 2 つの動的ユニットにおいて減衰振動性の過渡現象を励起します。これらのユニットの出力は、その後線形マッピングを介して 4 つの感覚チャネルに渡されます。ニューロンモデルまたは認識モデルの形式は、生成モデルと全く同じであることに注意してください。唯一の違いは、因果的状態が予測誤差によって駆動されることであり、これにより順方向接続（赤で示される）の必要性が生じます。条件付き不確実性（95

このシミュレーションは、感覚誘発過渡現象を再現すると見なすことができ、図の左側に示されている生成モデルのベイズ的逆問題解決に対応します。この文脈では、私たちが動的生成モデルを使用したため、逆問題解決はオンラインデコンボリューションに相当します。認識モデルの結合強度が自由エネルギーを最小化するように許容すると、対応する生成モデルのパラメータも暗黙的に推定しています。機械学習や信号処理では、これは盲目デコンボリューションとして知られています。この例を図 6 に示します。ここでは、同じ刺激を 8 回提示し、入力または最下層における予測誤差を、刺激周囲の全時間で合計して記録しました。パラメータの初期値は生成モデルと同じでした（図 5 で使用されたもの）。上部のパネルは、1 回目と最後の試行における刺激と予測された入力を画像形式で示しています。1 回目と 8 回目の予測は両方とも実際の入力とほぼ同一であることがわかります。これは、結合強度、すなわち（認識モデルにおける）パラメータの条件付きモードが、生成モデルによって使用されたのと同じ値で開始されたためです。

それにもかかわらず、パラメータに関するアンサンブル密度の自由エネルギーを最小化することで、認識モデルはこの刺激をより効率的に符号化でき、繰り返しの曝露とともに予測誤差が漸進的に抑制されます。この効果は、認識モデルがこれまでに見たことのない刺激を使用した場合にはるかに顕著です。この刺激は、生成モデルのパラメータに小さな乱数を加えることによって作成しました。最初の提示では、認識モデルは、すでに知っていることや経験したことに基づいて入力を知覚しようとします。この場合、期待される刺激の延長バージョンです。これは大きな予測誤差を生じます。8 回目の提示までには、パラメータの変化によ

り、入力をほぼ正確に認識および予測できるようになり、入力の繰り返しごとに予測誤差が深く抑制されます。予測されていない刺激では、予測誤差の抑制がより劇的であることに注意してください。これは、繰り返しの曝露中により多く学習されるためです。

9.2 反復抑制

この単純なシミュレーションは、自由エネルギー最小化スキーム、そして実際の脳の応答に普遍的かつ一般的な側面を示しています。それは反復抑制です。この現象は、刺激の繰り返し提示時に誘発反応が減少または抑制されることを指します。これは、EEG 研究におけるミスマッチ陰性電位（Na ☒ a ☒ ta ☒ nen, 2003）から、顔処理の fMRI 例（Henson et al., 2000 および図 7 参照）まで、多くの文脈で見られます。

私たちが焦点を当てる現象は、慣れた刺激や予測可能な刺激によって引き起こされる予測誤差と、予測不可能な刺激によって引き起こされる予測誤差との違いです。自由エネルギーの定式化の強い予測は、予測不可能または一貫性のない刺激が、慣れた刺激または一貫性のある刺激よりもはるかに大きな予測誤差を引き起こすということです。さらに、この相対的な抑制は、予測を伝える脳内の逆方向接続によって媒介されます。最後のセクションでは、予測可能および予測不可能な刺激を用いた視覚誘発反応の fMRI 研究（Harrison et al., in press）を用いて、この仮説の経験的検証を提示します。

10 ヒト脳における自由エネルギーの抑制

前のセクションで述べた自由エネルギーの扱いは、明らかに膨大な数の予測と実験につながります。私たちは、神経生理学、電気生理学、精神物理学、およびイメージング神経科学の文献から、これらの多くを他の論文でレビューしてきました（例：Friston and Price, 2001; Friston, 2003, 2005）。本論文では、予測可能および予測不可能な視覚刺激を用いて、逆方向接続の予測誤差抑制における役割を解明するために設計された、単純だが非常に示唆に富む研究に焦点を当てます。

10.1 実験デザインと方法

この実験は、最も単純に言えば、予測可能および予測不可能な刺激に対する視覚誘発反応を測定するものと捉えられます。我々は、初期（下位）視覚野における誘発反応は、予測不可能な刺激に比べて予測可能な刺激に対して減少すると仮説を立てました。

刺激は、一貫した（予測可能な）動きをするか、非一貫性の（予測不可能な）動きをする、疎なグリッド状の視覚的な点から構成されました。しかし、予測可能な刺激に対して単に反応の減少を示すだけでは、この減少が逆方向接続によって媒介されていると推論することはできません。

これを行うために、我々は視覚系における接続性の既知の解剖学的特徴を利用し、一貫性のある動きの効果が逆方向接続によって媒介されることを確実にしました。具体的には、縞状皮質（V1）の水平接続の範囲を超えて網膜局所的にマッピングされた反応を興奮させる疎な刺激を用いました。V1 ニューロンの古典的

受容野は約 1 視覚角度であるのに対し（図 8 参照）、V1 の水平接続は約 2 視覚角度をカバーします（Angelucci et al., 2002a,b）。我々が用いた刺激の分離は約 3 視覚角度でした。したがって、単一の点のいかなる成分運動も、他の点によって予測される可能性のあるものは、より大きな受容野を持つ高次視覚野（すなわち V2 以上）によってのみ「見られる」ことになります。このことは、V1 応答における予測可能性に起因する違いは、V2 以上の領域からの逆方向接続によって媒介されなければならないことを意味します。

非一貫性および全体的に一貫性のある疎な刺激は、通常のヒト被験者に約 1 秒ごとに提示され、その間、機能的磁気共鳴画像法（fMRI）を用いて血行動態反応が測定されました。データは、従来の統計的パラメトリックマッピングを用いて解析されました。これには、一貫性または非一貫性の刺激の発生を符号化する刺激関数を用いて誘発反応をモデル化し、これらを血行動態反応関数と畳み込み、一般線形モデルの回帰変数を作成する作業が含まれていました。一貫性刺激と非一貫性刺激の間の微分応答に関する推論は、各被験者からの適切なコントラストに基づいて、被験者間での単一標本 t 検定を用いて評価されました。

このランダム効果分析の結果を図 9 に示します。

10.2 結果

予測通り、V1 では予測可能な視覚刺激に対して、非一貫性の視覚刺激と比較して、視覚誘発反応が著しく減少しました。興味深いことに、これらの減少は V5 でも両側性に観察されました。単一被験者における V1、V2、V5 の血行動態活動の時間経過は上部のパネルに示されています。この図は、動かない対照刺激に対する推定応答も示しています。

ここでも、予測通り、受容野が複数の点を含むことができた最初のレベルである V2 領域で、予測不可能な刺激に対する応答の増強が見られました。これは、複数の点によって支えられたグローバルな動きを符号化する深層錐体細胞の活動を反映している可能性があります。V5 が予測誤差の減少を示したことは興味深い点です。V5 は一般的に V2 よりも視覚皮質において階層的に高い位置にあると考えられているにもかかわらずです。しかし、外側膝状体外経路は V1 と V2 をバイパスし、直接 V5 に情報を提供することができ、ある状況下では V5 が階層的に低い領域のように振る舞う可能性があります。これは、V1 と比較した V5 の短い潜時応答と一致しています（Nowak and Bullier, 1997 参照）。

まとめると、この fMRI 研究は、理論的分析からの私たちの予測、すなわち予測可能な刺激に対する誘発反応は、予測不可能な刺激に比べて小さいことを裏付けています。これは、感覚皮質がその膝状体入力の原因を推論するために自己組織化する際に引き起こされる予測誤差を、測定された応答が大部分反映していることと一致します。さらに、V1 ニューロン間の水平相互作用を排除するように刺激を慎重に設計することで、この予測誤差の抑制が高次皮質領域からの逆方向接続によって媒介されなければならないと推論することができます。これは、脳における生成モデルの階層的定式化と、自由エネルギー原理に従ったこれらのモデルの逆問題解決によってもたらされる再帰的ダイナミクスと一致しています。

11　結論

本論文では、生物学的システムの特性を、非生物学的な自己組織化および散逸システムとの関連で考察しました。生物学的システムは環境に作用し、相転移を避けるために選択的に環境をサンプリングすることができます。相転移は、不可逆的にその構造を変化させてしまうものです。この適応的な交換は、自由エネルギー最小化という観点から形式化でき、そこでは生物の行動とその内部構成の両方が自由エネルギーを最小化します。この自由エネルギーは、生物の構成によって符号化されるアンサンブル密度と、それが曝される感覚データの関数です。自由エネルギーの最小化は、行動依存的な感覚入力の変化と、内部変化によって暗示されるアンサンブル密度を通じて発生します。低い自由エネルギーを維持できないシステムは、環境との関係が変化するにつれて相転移に遭遇するでしょう。したがって、生物学的システムが自由エネルギーを最小化することは、十分ではないにしても、必要なことなのです。

この自由エネルギーは、熱力学的な自由エネルギーではなく、情報理論的な量という観点から定式化された自由エネルギーです。ここで議論されている自由エネルギー原理は、熱力学の結果ではなく、個体群のダイナミクスと選択から生じます。簡単に言えば、自由エネルギーが低いシステムは、自由エネルギーが高いシステムよりも選択されるでしょう。自由エネルギーは、生物の構造によって内包される生成モデルの特定に基づいています。このモデルを特定することで、システムが自由エネルギー原理に適合する場合にどのように変化するかを予測できます。脳にとって、もっともらしいモデルは階層的な動的システムであり、そこでは神経活動が環境状態の条件付きモードを符号化し、その結合がこれらの状態が進化する因果的な文脈を符号化します。感覚入力の原因を推論するためのこのモデルのベイズ的逆問題解決は、自由エネルギーを最小化すること、または簡略化された仮定の下では予測誤差の抑制の自然な結果です。私たちは、予測可能な刺激の文脈において、予測誤差の相対的な抑制が、実際に脳内の逆方向結合によって媒介されていることを示す、単純だが説得力のある実験で結論を締めくくりました。これは、自由エネルギー降下スキームによって予測された通りです。

本論文で提示されたアイデアは、深い歴史を持っています。ヘルムホルツ（1860年）によって記述されたニューロンエネルギーの概念に始まり、合成による分析（Neisser, 1967）のようなアイデアや、ベイズ的逆問題解決や予測符号化（例：Ballard et al., 1983; Mumford, 1992; Dayan et al., 1995; Rao and Ballard, 1998）のようなより最近の定式化を網羅しています。本論文の具体的な貢献は、行動と知覚の両方をカバーする自由エネルギー原理の一般的な定式化を提供することです。さらに、この定式化は、機械学習と統計物理学の構成要素を理論生物学の選択主義的なアイデアと結びつけるために使用できます。